
# Task 1: Mining Ressources
Progress & Plans

# Mining resources

- Development of robust linguistic resources to process different types and genres of texts
  - knowledge resources about **media personalities**: recognizing and resolving references to named-entities;
  - **sentiment lexicons and grammars**: detecting the polarity of opinions about media personalities
  - **annotated corpora**: training different text classifiers and evaluating classification procedures

REACTION

# Mining resources (ongoing activities)

- **POWER-PT01** - Political Ontology for Web Entity Retrieval
- **SentiLex-PT01** – Sentiment Lexicon for Portuguese
- **NomesLex-PT01** – Lexicon of person names *(new)*
- **Sentidioms** – (semi-)frozen sentiment constructions targeting human entities *(new)*
- **SentiCorpus-PT09** – Sentiment annotated corpus of user comments to political debates

REACTION

# Power

- Bootstrap phase
  - Data extracted from highly authoritative sources: government and national elections committee websites
- **Enrichment phase**
  - Extraction of alternative (media) names of politicians from **Sapo Voxx**
  - Matching Voxx names with Power politicians using a learned model (C4.5 in Weka)
  - Precision & Recall = 93.8 %
- **Statistics**
  - 3590 Politicians
  - 3043 Offices in Political Institutions
  - 74 Political Associations (18 of them are in fact coalitions)
  - 5959 Mandates
- **Deployment**
  - POWER is available as an RDF file and via SPARQL endpoint http://xldb.fc.ul.pt/wiki/POWER-PT01_in_English

REACTION

# Lexicon of Person Names

- **NomesLex-PT01 -** is a lexicon of person names made up of **2,027 first names** and **8,019 surnames**, and corresponding frequencies.
- Names were selected from the public list of teachers’ 2009 recruitment, published at the Portuguese Ministry of Education website.

  Most-frequent first names: **MARIA** (14374); **ANA** (7966); **CARLA** (2660); **SANDRA** (2300); **PAULA** (2021).

  Most-frequent surnames: **SILVA** (10170); **SANTOS** (6785); **FERREIRA** (5918); **PEREIRA** (5557); **OLIVEIRA** (4614).

REACTION

# Sentiment Lexicon

- Human Predicate Adjectives (6,321 adjective lemmas, and 25,406 inflected forms) – *under validation*.

- Human Predicate Nouns (966 lemmas)

  **agressividade**.PoS=N;GN=fs;TG=HUM;POL=-1;ANOT=MAN (aggressiveness)

  **altruísmo**.PoS=N;GN=ms;TG=HUM;POL=1;ANOT=MAN (altruism)

  **amnésia**.PoS=N;GN=fs;TG=HUM;POL=-1;ANOT=MAN (amnesia)

- Human Predicate Verbs (270 lemmas)

  **insultar**.PoS=V; TG=HUM:N0:N1;POL:N0=-1;POL:N1=0 (to insult)

  **mentir**.PoS=V; TG=HUM:N0;POL:N0=-1 (to lie)

  **derrotar**.PoS=V; TG=HUM:N0:N1;POL:N0=1;POL:N1=-1 (to defeat)

# Idiomatic expressions

- Human Idiomatic Expressions (472 canonical forms)

<N0> **<abandonar> o barco**. IDIOM; TG=HUM:N0; POL:N0=-1
(abandon ship)

<N0> **<acertar> na mosca**. IDIOM; TG=HUM:N0; POL:N0=1
(hit the bull´s eye)

<N0> **<agarrar> o touro pelos cornos**. IDIOM; TG=HUM:N0; POL:N0=1
(take the bull by the horns)

<N0> **<apunhalar>** <N1> **pelas costas**. IDIOM; TG=HUM:N0:N1; POL:N0=-1; POL:N1=0
(stab in the back)

# Publications

- ✓ Silvio Moreira, David Batista, Paula Carvalho, Francisco Couto, Mário J. Silva. «**POWER - Politics Ontology for Web Entity Retrieval**». ONTOSE 2011: 5th International Workshop on Ontology, Models, Conceptualization and Epistemology in Social, Artificial and Natural Systems, CAiSE 2011, London, United Kingdom, 20-24 June, 2011.

- ✓ Paula Carvalho, Luís Sarmento, Jorge Teixeira, Mário J. Silva. «**Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates**». The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, June 19-24, 2011.

REACTION

# Next steps (August 2011)

- POWER
  - Data enrichement
  - Data cleaning
- SentiCorpus-PT09 + SentiLex-PT02 + Sentidioms-PT01
  - Publicly available
- Analysis and (semi-automated) annotation of a collection of documents from industrial and social media, over a period of 6 months